EXHIBIT J

Trademark/Service Mark Application, Principal Register

Serial Number: 98046928
Filing Date: 06/16/2023

The table below presents the data as entered.

Input Field	Entered
SERIAL NUMBER	98046928
MARK INFORMATION	
*MARK	ANTHROPIC
STANDARD CHARACTERS	YES
USPTO-GENERATED IMAGE	YES
LITERAL ELEMENT	ANTHROPIC
MARK STATEMENT	The mark consists of standard characters, without claim to any particular font style, size, or color.
REGISTER	Principal
APPLICANT INFORMATION	
*OWNER OF MARK	Anthropic, PBC
*MAILING ADDRESS	548 Market St, PMB 90375
*CITY	San Francisco
*STATE (Required for U.S. applicants)	California
*COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
*ZIP/POSTAL CODE (Required for U.S. and certain international addresses)	94104
*EMAIL ADDRESS	XXXX
LEGAL ENTITY INFORMATION	
TYPE	benefit corporation
STATE/COUNTRY/REGION/JURISDICTION/U.S. TERRITORY WHERE LEGALLY ORGANIZED	Delaware
OWNER DOMICILE ADDRESS(NEW)	
*ADDRESS	XXXX
*CITY	XXXX
*STATE (Required for U.S. applicants)	XXXX
*COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	XXXX
*ZIP/POSTAL CODE (Required for U.S. and certain international addresses)	XXXX
GOODS AND/OR SERVICES AND BASIS INFORMATION	
INTERNATIONAL CLASS	

INTERNATIONAL CLASS	009
*IDENTIFICATION	Downloadable computer software in the nature of an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; downloadable software in the nature of a downloadable mobile application featuring an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; downloadable electronic publications in the nature of articles, research papers, reports, and newsletters in the fields of artificial intelligence, computer science, and computer software design and development
FILING BASIS	SECTION 1(b)

ATTORNEY INFORMATION

NAME	Betsy Wang Lee
ATTORNEY DOCKET NUMBER	48377-6000
ATTORNEY BAR MEMBERSHIP NUMBER	XXX
YEAR OF ADMISSION	XXXX
U.S. STATE/ COMMONWEALTH/ TERRITORY	XX
FIRM NAME	Orrick, Herrington & Sutcliffe LLP
INTERNAL ADDRESS	IP Prosecution Department
STREET	2050 Main Street, Suite 1100
CITY	Irvine
STATE	California
COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
ZIP/POSTAL CODE	92614-8255
PHONE	415-773-5700
FAX	949-567-6710
EMAIL ADDRESS	ipprosecutionsf@orrick.com
OTHER APPOINTED ATTORNEY	Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick

CORRESPONDENCE INFORMATION

NAME	Betsy Wang Lee
PRIMARY EMAIL ADDRESS FOR CORRESPONDENCE	ipprosecutionsf@orrick.com
SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES)	blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com

FEE INFORMATION

APPLICATION FILING OPTION	TEAS Standard
NUMBER OF CLASSES	1
APPLICATION FOR REGISTRATION PER CLASS	350

*TOTAL FEES DUE	350
*TOTAL FEES PAID	350
SIGNATURE INFORMATION	
SIGNATURE	/Ethan Forrest/
SIGNATORY'S NAME	Ethan Forrest
SIGNATORY'S POSITION	Associate General Counsel
SIGNATORY'S PHONE NUMBER	415-773-5700
DATE SIGNED	06/16/2023
SIGNATURE METHOD	Sent to third party for signature

Trademark/Service Mark Application, Principal Register

Serial Number: 98046928
Filing Date: 06/16/2023

To the Commissioner for Trademarks:

MARK: ANTHROPIC (Standard Characters, see mark)
The literal element of the mark consists of ANTHROPIC. The mark consists of standard characters, without claim to any particular font style, size, or color.
The applicant, Anthropic, PBC, a benefit corporation legally organized under the laws of Delaware, having an address of
   548 Market St, PMB 90375
   San Francisco, California 94104
   United States
   XXXX

and a domicile address of

   XXXX
   XXXX, XXXX XXXX
   XXXX

requests registration of the trademark/service mark identified above in the United States Patent and Trademark Office on the Principal Register established by the Act of July 5, 1946 (15 U.S.C. Section 1051 et seq.), as amended, for the following:

International Class 009: Downloadable computer software in the nature of an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; downloadable software in the nature of a downloadable mobile application featuring an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; downloadable electronic publications in the nature of articles, research papers, reports, and newsletters in the fields of artificial intelligence, computer science, and computer software design and development
Intent to Use: The applicant has a bona fide intention, and is entitled, to use the mark in commerce on or in connection with the identified goods/services.

The owner's/holder's proposed attorney information: Betsy Wang Lee. Other appointed attorneys are Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick. Betsy Wang Lee of Orrick, Herrington & Sutcliffe LLP, is a member of the XX bar, admitted to the bar in XXXX, bar membership no. XXX, and the attorney(s) is located at
   IP Prosecution Department
   2050 Main Street, Suite 1100
   Irvine, California 92614-8255
   United States
   415-773-5700(phone)
   949-567-6710(fax)
   ipprosecutionsf@orrick.com
The docket/reference number is 48377-6000.
Betsy Wang Lee submitted the following statement: The attorney of record is an active member in good standing of the bar of the highest court of a U.S. state, the District of Columbia, or any U.S. Commonwealth or territory.
The applicant's current Correspondence Information:
   Betsy Wang Lee

   PRIMARY EMAIL FOR CORRESPONDENCE: ipprosecutionsf@orrick.com
   SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES): blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com

Requirement for Email and Electronic Filing: I understand that a valid email address must be maintained by the applicant owner/holder and the applicant owner's/holder's attorney, if appointed, and that all official trademark correspondence must be submitted via the Trademark Electronic Application System (TEAS).

A fee payment in the amount of $350 has been submitted with the application, representing payment for 1 class(es).

<div align="center">Declaration</div>

☑ **Basis:**
If the applicant is filing the application based on use in commerce under 15 U.S.C. § 1051(a):

- The signatory believes that the applicant is the owner of the trademark/service mark sought to be registered;
- The mark is in use in commerce and was in use in commerce as of the filing date of the application on or in connection with the goods/services in the application;
- The specimen(s) shows the mark as used on or in connection with the goods/services in the application and was used on or in connection with the goods/services in the application as of the application filing date; and
- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

And/Or
If the applicant is filing the application based on an intent to use the mark in commerce under 15 U.S.C. § 1051(b), § 1126(d), and/or § 1126(e):

- The signatory believes that the applicant is entitled to use the mark in commerce;
- The applicant has a bona fide intention to use the mark in commerce and had a bona fide intention to use the mark in commerce as of the application filing date on or in connection with the goods/services in the application; and
- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

☑ To the best of the signatory's knowledge and belief, no other persons, except, if applicable, concurrent users, have the right to use the mark in commerce, either in the identical form or in such near resemblance as to be likely, when used on or in connection with the goods/services of such other persons, to cause confusion or mistake, or to deceive.

☑ To the best of the signatory's knowledge, information, and belief, formed after an inquiry reasonable under the circumstances, the allegations and other factual contentions made above have evidentiary support.

☑ The signatory being warned that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001, and that such willful false statements and the like may jeopardize the validity of the application or submission or any registration resulting therefrom, declares that all statements made of his/her own knowledge are true and all statements made on information and belief are believed to be true.

Declaration Signature

Signature: /Ethan Forrest/   Date: 06/16/2023
Signatory's Name: Ethan Forrest
Signatory's Position: Associate General Counsel
Signatory's Phone Number: 415-773-5700
Signature method: Sent to third party for signature
Payment Sale Number: 98046928
Payment Accounting Date: 06/16/2023

Serial Number: 98046928
Internet Transmission Date: Fri Jun 16 17:18:04 ET 2023
TEAS Stamp: USPTO/BAS-XX.XX.XXX.XX-20230616171804708
263-98046928-860fab9427de129e3656ed9d64f
b4abbc05b851974b1b1297c438b4d7372eaa9b-D
A-18047801-20230615150743495092

ANTHROPIC

Trademark/Service Mark Application, Principal Register

Serial Number: 98046931
Filing Date: 06/16/2023

The table below presents the data as entered.

Input Field	Entered
SERIAL NUMBER	98046931
MARK INFORMATION	
*****MARK**	ANTHROPIC
STANDARD CHARACTERS	YES
USPTO-GENERATED IMAGE	YES
LITERAL ELEMENT	ANTHROPIC
MARK STATEMENT	The mark consists of standard characters, without claim to any particular font style, size, or color.
REGISTER	Principal
APPLICANT INFORMATION	
*****OWNER OF MARK**	Anthropic, PBC
*****MAILING ADDRESS**	548 Market St, PMB 90375
*****CITY**	San Francisco
*****STATE** (Required for U.S. applicants)	California
*****COUNTRY/REGION/JURISDICTION/U.S. TERRITORY**	United States
*****ZIP/POSTAL CODE** (Required for U.S. and certain international addresses)	94104
*****EMAIL ADDRESS**	XXXX
LEGAL ENTITY INFORMATION	
TYPE	benefit corporation
STATE/COUNTRY/REGION/JURISDICTION/U.S. TERRITORY WHERE LEGALLY ORGANIZED	Delaware
OWNER DOMICILE ADDRESS(NEW)	
*****ADDRESS**	XXXX
*****CITY**	XXXX
*****STATE** (Required for U.S. applicants)	XXXX
*****COUNTRY/REGION/JURISDICTION/U.S. TERRITORY**	XXXX
*****ZIP/POSTAL CODE** (Required for U.S. and certain international addresses)	XXXX
GOODS AND/OR SERVICES AND BASIS INFORMATION	
INTERNATIONAL CLASS	

INTERNATIONAL CLASS	035
*IDENTIFICATION	Public policy research; public policy consultation; public policy research related to the societal implications of artificial intelligence; promoting public awareness of scientific research and development, artificial intelligence, the ethical aspects of artificial intelligence, computer science, and computer software design and development; providing business consulting services in the fields of artificial intelligence, computer science, and computer software design and development
FILING BASIS	SECTION 1(a)
FIRST USE ANYWHERE DATE	At least as early as 12/01/2021
FIRST USE IN COMMERCE DATE	At least as early as 12/01/2021
SPECIMEN FILE NAME(S)	
ORIGINAL PDF FILE	SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_Specimen_1.pdf
CONVERTED PDF FILE(S) (4 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0003.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0004.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0005.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0006.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_Specimen_2.pdf
CONVERTED PDF FILE(S) (3 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0007.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0008.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0009.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_Specimen_3.pdf
CONVERTED PDF FILE(S) (2 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0010.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0011.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_Specimen_4.pdf
CONVERTED PDF FILE(S) (2 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0012.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046931\xml1\ APP0013.JPG
SPECIMEN DESCRIPTION	screenshots of promotional materials from Applicant's website and research article showing the mark being used in

	connection with the services listed in the application
WEBPAGE URL	https://www.anthropic.com/company
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/research
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/index/predictability-and-surprise-in-large-generative-models
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://cdn2.assets-servd.host/anthropic-website/production/images/Anthropic_PredictabilityAndSurprise.pdf
WEBPAGE DATE OF ACCESS	06/15/2023
ATTORNEY INFORMATION	
NAME	Betsy Wang Lee
ATTORNEY DOCKET NUMBER	48377-6000
ATTORNEY BAR MEMBERSHIP NUMBER	XXX
YEAR OF ADMISSION	XXXX
U.S. STATE/ COMMONWEALTH/ TERRITORY	XX
FIRM NAME	Orrick, Herrington & Sutcliffe LLP
INTERNAL ADDRESS	IP Prosecution Department
STREET	2050 Main Street, Suite 1100
CITY	Irvine
STATE	California
COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
ZIP/POSTAL CODE	92614-8255
PHONE	415-773-5700
FAX	949-567-6710
EMAIL ADDRESS	ipprosecutionsf@orrick.com
OTHER APPOINTED ATTORNEY	Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick
CORRESPONDENCE INFORMATION	
NAME	Betsy Wang Lee
PRIMARY EMAIL ADDRESS FOR CORRESPONDENCE	ipprosecutionsf@orrick.com
SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES)	blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com
FEE INFORMATION	
APPLICATION FILING OPTION	TEAS Standard
NUMBER OF CLASSES	1
APPLICATION FOR REGISTRATION PER CLASS	350

*TOTAL FEES DUE	350
*TOTAL FEES PAID	350
SIGNATURE INFORMATION	
SIGNATURE	/Ethan Forrest/
SIGNATORY'S NAME	Ethan Forrest
SIGNATORY'S POSITION	Associate General Counsel
SIGNATORY'S PHONE NUMBER	415-773-5700
DATE SIGNED	06/16/2023
SIGNATURE METHOD	Sent to third party for signature

Trademark/Service Mark Application, Principal Register

Serial Number: 98046931
Filing Date: 06/16/2023

To the Commissioner for Trademarks:

MARK: ANTHROPIC (Standard Characters, see mark)
The literal element of the mark consists of ANTHROPIC. The mark consists of standard characters, without claim to any particular font style, size, or color.
The applicant, Anthropic, PBC, a benefit corporation legally organized under the laws of Delaware, having an address of

    548 Market St, PMB 90375
    San Francisco, California 94104
    United States
    XXXX

and a domicile address of

    XXXX
    XXXX, XXXX XXXX
    XXXX

requests registration of the trademark/service mark identified above in the United States Patent and Trademark Office on the Principal Register established by the Act of July 5, 1946 (15 U.S.C. Section 1051 et seq.), as amended, for the following:

International Class 035: Public policy research; public policy consultation; public policy research related to the societal implications of artificial intelligence; promoting public awareness of scientific research and development, artificial intelligence, the ethical aspects of artificial intelligence, computer science, and computer software design and development; providing business consulting services in the fields of artificial intelligence, computer science, and computer software design and development

In International Class 035, the mark was first used by the applicant or the applicant's related company or licensee or predecessor in interest at least as early as 12/01/2021, and first used in commerce at least as early as 12/01/2021, and is now in use in such commerce. The applicant is submitting one(or more) specimen(s) showing the mark as used in commerce on or in connection with any item in the class of listed goods/services, consisting of a(n) screenshots of promotional materials from Applicant's website and research article showing the mark being used in connection with the services listed in the application.

Original PDF file:
SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_ Specimen_1.pdf
Converted PDF file(s) (4 pages)
Specimen File1
Specimen File2
Specimen File3
Specimen File4
Original PDF file:
SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_ Specimen_2.pdf
Converted PDF file(s) (3 pages)
Specimen File1
Specimen File2
Specimen File3
Original PDF file:
SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_ Specimen_3.pdf
Converted PDF file(s) (2 pages)
Specimen File1
Specimen File2
Original PDF file:

SPE0-389810518-2023061515 2214759521_._ANTHROPIC_-_ Specimen_4.pdf
Converted PDF file(s) (2 pages)
Specimen File1
Specimen File2
Webpage URL: https://www.anthropic.com/company
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/research
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/index/predictability-and-surprise-in-large-generative-models
Webpage Date of Access: 06/15/2023
Webpage URL: https://cdn2.assets-servd.host/anthropic-website/production/images/Anthropic_PredictabilityAndSurprise.pdf
Webpage Date of Access: 06/15/2023

The owner's/holder's proposed attorney information: Betsy Wang Lee. Other appointed attorneys are Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick. Betsy Wang Lee of Orrick, Herrington & Sutcliffe LLP, is a member of the XX bar, admitted to the bar in XXXX, bar membership no. XXX, and the attorney(s) is located at

IP Prosecution Department
2050 Main Street, Suite 1100
Irvine, California 92614-8255
United States
415-773-5700(phone)
949-567-6710(fax)
ipprosecutionsf@orrick.com

The docket/reference number is 48377-6000.
Betsy Wang Lee submitted the following statement: The attorney of record is an active member in good standing of the bar of the highest court of a U.S. state, the District of Columbia, or any U.S. Commonwealth or territory.
The applicant's current Correspondence Information:

Betsy Wang Lee

PRIMARY EMAIL FOR CORRESPONDENCE: ipprosecutionsf@orrick.com
SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES): blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com

Requirement for Email and Electronic Filing: I understand that a valid email address must be maintained by the applicant owner/holder and the applicant owner's/holder's attorney, if appointed, and that all official trademark correspondence must be submitted via the Trademark Electronic Application System (TEAS).
A fee payment in the amount of $350 has been submitted with the application, representing payment for 1 class(es).

<div align="center">

Declaration

</div>

☑ **Basis:**
  If the applicant is filing the application based on use in commerce under 15 U.S.C. § 1051(a):

  - The signatory believes that the applicant is the owner of the trademark/service mark sought to be registered;
  - The mark is in use in commerce and was in use in commerce as of the filing date of the application on or in connection with the goods/services in the application;
  - The specimen(s) shows the mark as used on or in connection with the goods/services in the application and was used on or in connection with the goods/services in the application as of the application filing date; and
  - To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

  And/Or
  If the applicant is filing the application based on an intent to use the mark in commerce under 15 U.S.C. § 1051(b), § 1126(d), and/or § 1126(e):

  - The signatory believes that the applicant is entitled to use the mark in commerce;
  - The applicant has a bona fide intention to use the mark in commerce and had a bona fide intention to use the mark in commerce as of the application filing date on or in connection with the goods/services in the application; and
  - To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

☑ To the best of the signatory's knowledge and belief, no other persons, except, if applicable, concurrent users, have the right to use the mark in commerce, either in the identical form or in such near resemblance as to be likely, when used on or in connection with the goods/services of such other persons, to cause confusion or mistake, or to deceive.

☑ To the best of the signatory's knowledge, information, and belief, formed after an inquiry reasonable under the circumstances, the allegations and other factual contentions made above have evidentiary support.

☑ The signatory being warned that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001, and that such willful false statements and the like may jeopardize the validity of the application or submission or any registration resulting therefrom, declares that all statements made of his/her own knowledge are true and all statements made on information and belief are believed to be true.

Declaration Signature

Signature: /Ethan Forrest/   Date: 06/16/2023
Signatory's Name: Ethan Forrest
Signatory's Position: Associate General Counsel
Signatory's Phone Number: 415-773-5700
Signature method: Sent to third party for signature
Payment Sale Number: 98046931
Payment Accounting Date: 06/16/2023

Serial Number: 98046931
Internet Transmission Date: Fri Jun 16 17:18:27 ET 2023
TEAS Stamp: USPTO/BAS-XX.XX.XXX.XX-20230616171827632
471-98046931-86097d4146476f507d6f6ce646f
5700c2bf89353cd5fae70acb994b047e7896f-DA
-18277252-20230615174700583929

ANTHROPIC

ANTHROP\C

Making AI systems you can rely on

Anthropic is an AI safety and research company. We build reliable, interpretable, and steerable AI systems.

[Join Us]

Our Purpose

We believe AI will have a vast impact on the world. Anthropic is dedicated to building systems that people can rely on and generating research about the opportunities and risks of AI.

We Build Safer Systems

We aim to build frontier AI systems that are reliable, interpretable, and steerable. We conduct frontier research, develop and apply a variety of safety techniques, and deploy the resulting systems via a set of partnerships and products.

Safety Is a Science

We treat AI safety as a systematic science, conducting research, applying it to our products, feeding those insights back into our research, and regularly sharing what we learn with the world along the way.

Interdisciplinary

Anthropic is a collaborative team of researchers, engineers, policy experts, business leaders and operators, who bring our experience from many different domains to our work.

AI Companies are One Piece of a Big Puzzle

AI has the potential to fundamentally change how the world works. We view ourselves as just one piece of this evolving puzzle. We collaborate with civil society, government, academia, nonprofits and

industry to promote safety industry-wide.

The Team

We're a team of researchers, engineers, policy experts and operational leaders, with experience spanning a variety of disciplines, all working together to build reliable and understandable AI systems.

Research

We conduct frontier AI research across a variety of modalities, and explore novel and emerging safety research areas from interpretability to RL from human feedback to policy and societal impacts analysis.

Policy

We think about the impacts of our work and strive to communicate what we're seeing at the frontier to policymakers and civil society in the US and abroad to help promote safe and reliable AI.

Product

We translate our research into tangible, practical tools like Claude that benefit businesses, nonprofits and civil society groups and their clients and people around the globe.

Operations

Our people, finance, legal, and recruiting teams are the human engines that make Anthropic go. We've had previous careers at NASA, startups, and the armed forces and our diverse experiences help make Anthropic a great place to work (and we love plants!).

16

Our Values

01
Here for the mission

Anthropic exists for our mission: to ensure transformative AI helps people and society flourish. Progress this decade may be rapid, and we expect increasingly capable systems to pose novel challenges. We pursue our mission by building frontier systems, studying their behaviors, working to responsibly deploy them, and regularly sharing our safety insights. We collaborate with other projects and stakeholders seeking a similar outcome.

04
Do the simple thing that works

We celebrate trying the simple thing before the clever, novel thing. We embrace pragmatism - sensible, practical approaches that acknowledge tradeoffs. We love empiricism - finding out what actually works by trying it - and apply this to our research, our engineering and our collaboration. We aim to be open about what we understand and what we don't.

02
Unusually high trust

Our company is an unusually high trust environment: we assume good faith, disagree kindly, and prioritize honesty. We expect emotional maturity and intellectual openness. At its best, our trust enables us to make better decisions as an organization than any one of us could as individuals.

03
One big team

Collaboration is central to our work, culture, and value proposition. While we have many teams at Anthropic, we feel the broader sense in which we are all on the same team working together towards the mission. Leadership sets the strategy, with broad input from everyone, and trusts each piece of the organization to pursue these goals in their unique style. Individuals commonly contribute to work across many different areas.

Company News

See All

Charting a Path to AI Accountability

Jun 13, 2023 · 4 min read

Anthropic Raises $450 Million in Series C Funding to Scale Reliable AI Products

May 23, 2023 · 2 min read

Zoom Partnership and Investment in Anthropic

May 16, 2023 · 56 sec read

Want to help us build the future of safe AI?

Join Us

A\

Product	Press Inquiries	Terms of Service	© 2023 Anthropic PBC
Research	Support	Privacy Policy	
Index	Twitter	Responsible Disclosure Policy	
Company	LinkedIn	Compliance	
News			
Careers			

ANTHROP\C

Product   Research   Company   News   Careers

Make safe AI systems Deploy them reliably

We develop large-scale AI systems so that we can study
their safety properties at the technological frontier,
where new problems are most likely to arise. We use
these insights to create safer, steerable, and more reliable
models, and to generate systems that we deploy
externally, like Claude.

See Research Principles

Alignment

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Our second AI alignment paper, exploring
how to train a general language assistant to
be helpful, but without providing harmful
advice or exhibiting bad behaviors.

Apr 12, 2022

Search　　　　　　　　　　　　　　　　　　　　　　　Q　　　Refine ≡　　View ≡

Circuits Updates — May 2023	May 24, 2023
Interpretability Dreams	May 24, 2023
Distributed Representations: Composition & Superposition	May 4, 2023
Privileged Bases in the Transformer Residual Stream	Mar 16, 2023
The Capacity for Moral Self-Correction in Large Language Models	Feb 15, 2023
Superposition, Memorization, and Double Descent	Jan 5, 2023
Discovering Language Model Behaviors with Model-Written Evaluations	Dec 19, 2022
Constitutional AI: Harmlessness from AI Feedback	Dec 15, 2022
Measuring Progress on Scalable Oversight for Large Language Models	Nov 4, 2022
Toy Models of Superposition	Sep 14, 2022
Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned	Aug 22, 2022
Language Models (Mostly) Know What They Know	Jul 11, 2022
Softmax Linear Units	Jun 17, 2022
Scaling Laws and Interpretability of Learning from Repeated Data	May 21, 2022
In-context Learning and Induction Heads	Mar 8, 2022

Predictability and Surprise in Large Generative Models   Feb 15, 2022

A Mathematical Framework for Transformer Circuits   Dec 22, 2021

A General Language Assistant as a Laboratory for Alignment   Dec 1, 2021

A

Product	Press Inquiries	Terms of Service	
Research	Support	Privacy Policy	
Index	Twitter	Responsible Disclosure Policy	
Company	LinkedIn		
News		Compliance	
Careers			

ANTHROP\C   Product   Research   Company   News   Careers

(Research) (Societal Impact)

Predictability and Surprise in Large Generative Models

Feb 15, 2022

Read Paper

Abstract

Large-scale pre-training has recently emerged as a technique for creating capable, general purpose, generative models such as GPT-3, Megatron-Turing NLG, Gopher, and many others. In this paper, we highlight a counterintuitive property of such models and discuss the policy implications of this property. Namely, these generative models have an unusual combination of predictable loss on a broad training distribution (as embodied in their "scaling laws"), and unpredictable specific capabilities, inputs, and outputs. We believe that the high-level predictability and appearance of useful capabilities drives rapid development of such models, while the unpredictable qualities make it difficult to anticipate the consequences of model deployment. We go through examples of how this combination can lead to socially harmful behavior with examples from the literature and real world observations, and we also perform two novel experiments to illustrate our point about harms from unpredictability. Furthermore, we analyze how these conflicting properties combine to give model developers various motivations for deploying these models, and challenges that can hinder deployment. We conclude with a list of possible interventions the AI community may take to increase the chance of these models having a beneficial impact. We intend this paper to be useful to policymakers who want to understand and regulate AI systems, technologists who care about the potential policy impact of their work, and academics who want to analyze, critique, and potentially develop large generative models.

Policy Memo

Predictability and Surprise Memo

f in y

Related

Research

The Capacity for Moral Self-Correction in Large Language Models

Feb 15, 2023

Research

Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned

Aug 22, 2022

Aᛋ

Product

Research

Index

Company

News

Careers

Press Inquiries

Support

Twitter

LinkedIn

Terms of Service

Privacy Policy

Responsible Disclosure Policy

Compliance

© 2023 Anthropic PBC
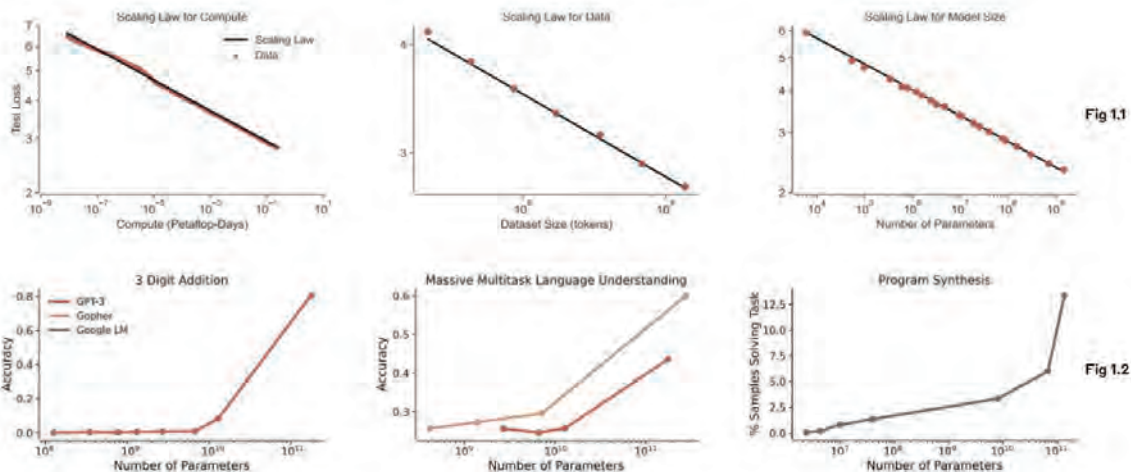
ANTHROP\C

Predictability and Surprise in Large Generative Models

In recent years, AI researchers have built far more general and useful AI systems. The best current examples are models that can process and generate text (e.g. OpenAI's *GPT-3*, Microsoft & NVIDIA's *Megatron-Turing NLG*, and DeepMind's *Chinchilla*). Unlike previous AI systems designed to perform a single task, these models can be used in a variety of applications, without being explicitly trained for those purposes.

These models share an unusual combination of interrelated characteristics. In some respects they are reliably predictable, while in other ways, they are quite unpredictable. Aggregate model performance follows a *predictable* trend in relation to the resources expended on training. **By scaling up the size of models, the computing power (compute) used to train them, and the amount of data they're trained on (in the correct proportions), models demonstrate improved general performance in a predictable manner (a trend referred to as a "scaling law," Fig 1.1).**

On the other hand, individual capabilities and outputs can't be predicted ahead of time — developers can't tell you precisely what new behaviors will emerge as they scale up models. For example, the ability to complete a specific task can sometimes emerge abruptly as developers increase the size of a model (Fig 1.2). The unpredictable nature of these models makes it difficult to fully account for the consequences of their development and deployment, demonstrating the importance of empirical safety research.
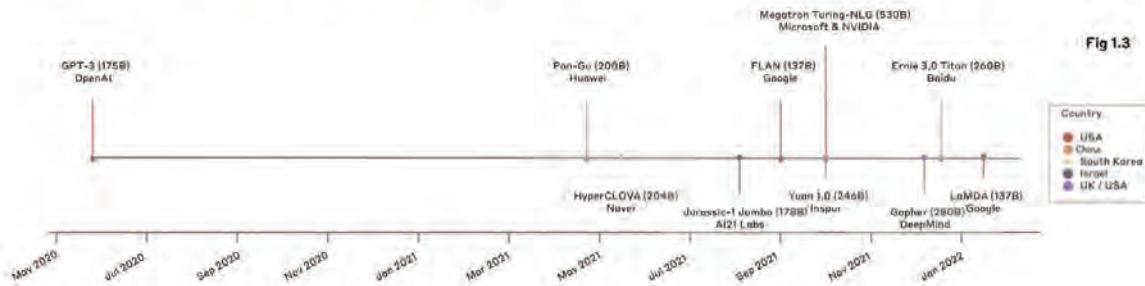

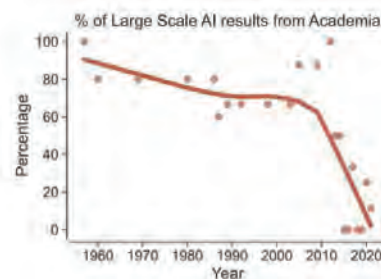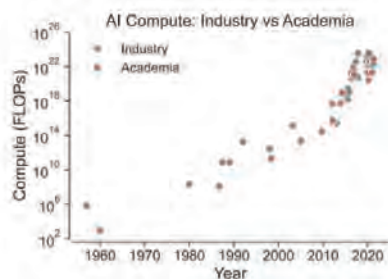Fig 1.1

Fig 1.2

24

ANTHROP\C

Despite these risks, more organizations from around the world are developing these systems, as a significant amount of the uncertainty inherent to AI development has been reduced (Fig 1.3).

Organizations (or countries) may also be motivated by the scientific potential these models create for novel AI research or the prestige associated with being on the technological frontier.

Fig 1.3

Due to the expense involved in building these models and the technical talent required to engineer them, **private sector organizations from a number of countries are the ones building these models – not academia, civil society, or public sector organizations.** Over the past decade, we've seen a dramatic shift in AI R&D as more computationally-intensive research is conducted by private sector actors, while academia increasingly lags in its ability to build or investigate models at the frontier (Fig 1.4).

AI Compute: Industry vs Academia

% of Large Scale AI results from Academia

Fig 1.4

Based on the distinguishing features of these models, and the economic motivations for their development and deployment, we predict large generative models will increasingly be developed and deployed despite their potential for harm. We think there are some policy interventions available that can increase the chance of these models being developed and deployed in positive ways:

- Reduce compute asymmetries between the private sector and academia
- Improve the technical tools publicly available for model evaluation
- Increase our understanding of abrupt jumps in capabilities
- Improve shared knowledge of how to "red team" models
- Explore and prototype novel governance structures and government interventions

25

Trademark/Service Mark Application, Principal Register

Serial Number: 98046932
Filing Date: 06/16/2023

The table below presents the data as entered.

Input Field	Entered
SERIAL NUMBER	98046932
MARK INFORMATION	
***MARK**	ANTHROPIC
STANDARD CHARACTERS	YES
USPTO-GENERATED IMAGE	YES
LITERAL ELEMENT	ANTHROPIC
MARK STATEMENT	The mark consists of standard characters, without claim to any particular font style, size, or color.
REGISTER	Principal
APPLICANT INFORMATION	
***OWNER OF MARK**	Anthropic, PBC
***MAILING ADDRESS**	548 Market St, PMB 90375
***CITY**	San Francisco
***STATE** (Required for U.S. applicants)	California
***COUNTRY/REGION/JURISDICTION/U.S. TERRITORY**	United States
***ZIP/POSTAL CODE** (Required for U.S. and certain international addresses)	94104
***EMAIL ADDRESS**	XXXX
LEGAL ENTITY INFORMATION	
TYPE	benefit corporation
STATE/COUNTRY/REGION/JURISDICTION/U.S. TERRITORY WHERE LEGALLY ORGANIZED	Delaware
OWNER DOMICILE ADDRESS(NEW)	
***ADDRESS**	XXXX
***CITY**	XXXX
***STATE** (Required for U.S. applicants)	XXXX
***COUNTRY/REGION/JURISDICTION/U.S. TERRITORY**	XXXX
***ZIP/POSTAL CODE** (Required for U.S. and certain international addresses)	XXXX
GOODS AND/OR SERVICES AND BASIS INFORMATION	
INTERNATIONAL CLASS	

26

INTERNATIONAL CLASS	041
*IDENTIFICATION	Educational services, namely, organizing and conducting live and online training in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; providing online publications in the nature of articles, research papers, reports, and newsletters in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; providing a website featuring non-downloadable publications in the nature of articles, research papers, reports, and newsletters in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; educational research in the fields of artificial intelligence, computer science, and computer software design and development
FILING BASIS	SECTION 1(a)
FIRST USE ANYWHERE DATE	At least as early as 02/07/2023
FIRST USE IN COMMERCE DATE	At least as early as 02/07/2023
SPECIMEN FILE NAME(S)	
ORIGINAL PDF FILE	SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_Specimen_1.pdf
CONVERTED PDF FILE(S) (4 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0003.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0004.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0005.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0006.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_Specimen_2.pdf
CONVERTED PDF FILE(S) (3 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0007.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0008.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0009.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_Specimen_5.pdf
CONVERTED PDF FILE(S) (2 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0010.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0011.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_Specimen_4.pdf
CONVERTED PDF FILE(S) (2 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0012.JPG

	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046932\xml1\ APP0013.JPG
SPECIMEN DESCRIPTION	screenshots of promotional materials from Applicant's website and research article showing the mark being used in connection with the services listed in the application
WEBPAGE URL	https://www.anthropic.com/company
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/research
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/index/discovering-language-model-behaviors-with-model-written-evaluations
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://cdn2.assets-servd.host/anthropic-website/production/images/Anthropic_PredictabilityAndSurprise.pdf
WEBPAGE DATE OF ACCESS	06/15/2023
ATTORNEY INFORMATION	
NAME	Betsy Wang Lee
ATTORNEY DOCKET NUMBER	48377-6000
ATTORNEY BAR MEMBERSHIP NUMBER	XXX
YEAR OF ADMISSION	XXXX
U.S. STATE/ COMMONWEALTH/ TERRITORY	XX
FIRM NAME	Orrick, Herrington & Sutcliffe LLP
INTERNAL ADDRESS	IP Prosecution Department
STREET	2050 Main Street, Suite 1100
CITY	Irvine
STATE	California
COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
ZIP/POSTAL CODE	92614-8255
PHONE	415-773-5700
FAX	949-567-6710
EMAIL ADDRESS	ipprosecutionsf@orrick.com
OTHER APPOINTED ATTORNEY	Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick
CORRESPONDENCE INFORMATION	
NAME	Betsy Wang Lee
PRIMARY EMAIL ADDRESS FOR CORRESPONDENCE	ipprosecutionsf@orrick.com
SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES)	blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com
FEE INFORMATION	

APPLICATION FILING OPTION	TEAS Standard
NUMBER OF CLASSES	1
APPLICATION FOR REGISTRATION PER CLASS	350
*TOTAL FEES DUE	350
*TOTAL FEES PAID	350
SIGNATURE INFORMATION	
SIGNATURE	/Ethan Forrest/
SIGNATORY'S NAME	Ethan Forrest
SIGNATORY'S POSITION	Associate General Counsel
SIGNATORY'S PHONE NUMBER	415-773-5700
DATE SIGNED	06/16/2023
SIGNATURE METHOD	Sent to third party for signature

Trademark/Service Mark Application, Principal Register

Serial Number: 98046932
Filing Date: 06/16/2023

To the Commissioner for Trademarks:

MARK: ANTHROPIC (Standard Characters, see mark)
The literal element of the mark consists of ANTHROPIC. The mark consists of standard characters, without claim to any particular font style, size, or color.
The applicant, Anthropic, PBC, a benefit corporation legally organized under the laws of Delaware, having an address of

    548 Market St, PMB 90375
    San Francisco, California 94104
    United States
    XXXX

and a domicile address of

    XXXX
    XXXX, XXXX XXXX
    XXXX

requests registration of the trademark/service mark identified above in the United States Patent and Trademark Office on the Principal Register established by the Act of July 5, 1946 (15 U.S.C. Section 1051 et seq.), as amended, for the following:

International Class 041: Educational services, namely, organizing and conducting live and online training in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; providing online publications in the nature of articles, research papers, reports, and newsletters in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; providing a website featuring non-downloadable publications in the nature of articles, research papers, reports, and newsletters in the fields of scientific research and development, artificial intelligence, computer science, and computer software design and development; educational research in the fields of artificial intelligence, computer science, and computer software design and development

In International Class 041, the mark was first used by the applicant or the applicant's related company or licensee or predecessor in interest at least as early as 02/07/2023, and first used in commerce at least as early as 02/07/2023, and is now in use in such commerce. The applicant is submitting one(or more) specimen(s) showing the mark as used in commerce on or in connection with any item in the class of listed goods/services, consisting of a(n) screenshots of promotional materials from Applicant's website and research article showing the mark being used in connection with the services listed in the application.

Original PDF file:
SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_ Specimen_1.pdf
Converted PDF file(s) (4 pages)
Specimen File1
Specimen File2
Specimen File3
Specimen File4
Original PDF file:
SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_ Specimen_2.pdf
Converted PDF file(s) (3 pages)
Specimen File1
Specimen File2
Specimen File3
Original PDF file:
SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_Specimen_5.pdf
Converted PDF file(s) (2 pages)
Specimen File1

_____
Specimen File2
Original PDF file:
SPE0-389810518-2023061515 3124486186_._ANTHROPIC_-_ Specimen_4.pdf
Converted PDF file(s) (2 pages)
Specimen File1
Specimen File2
Webpage URL: https://www.anthropic.com/company
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/research
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/index/discovering-language-model-behaviors-with-model-written-evaluations
Webpage Date of Access: 06/15/2023
Webpage URL: https://cdn2.assets-servd.host/anthropic-website/production/images/Anthropic_PredictabilityAndSurprise.pdf
Webpage Date of Access: 06/15/2023

The owner's/holder's proposed attorney information: Betsy Wang Lee. Other appointed attorneys are Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick. Betsy Wang Lee of Orrick, Herrington & Sutcliffe LLP, is a member of the XX bar, admitted to the bar in XXXX, bar membership no. XXX, and the attorney(s) is located at

IP Prosecution Department
2050 Main Street, Suite 1100
Irvine, California 92614-8255
United States
415-773-5700(phone)
949-567-6710(fax)
ipprosecutionsf@orrick.com
The docket/reference number is 48377-6000.
Betsy Wang Lee submitted the following statement: The attorney of record is an active member in good standing of the bar of the highest court of a U.S. state, the District of Columbia, or any U.S. Commonwealth or territory.
The applicant's current Correspondence Information:

Betsy Wang Lee
 PRIMARY EMAIL FOR CORRESPONDENCE: ipprosecutionsf@orrick.com
 SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES): blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com

Requirement for Email and Electronic Filing: I understand that a valid email address must be maintained by the applicant owner/holder and the applicant owner's/holder's attorney, if appointed, and that all official trademark correspondence must be submitted via the Trademark Electronic Application System (TEAS).
A fee payment in the amount of $350 has been submitted with the application, representing payment for 1 class(es).

<div align="center">

Declaration

</div>

☑ **Basis:**
If the applicant is filing the application based on use in commerce under 15 U.S.C. § 1051(a):

- The signatory believes that the applicant is the owner of the trademark/service mark sought to be registered;
- The mark is in use in commerce and was in use in commerce as of the filing date of the application on or in connection with the goods/services in the application;
- The specimen(s) shows the mark as used on or in connection with the goods/services in the application and was used on or in connection with the goods/services in the application as of the application filing date; and
- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

And/Or
If the applicant is filing the application based on an intent to use the mark in commerce under 15 U.S.C. § 1051(b), § 1126(d), and/or § 1126(e):

- The signatory believes that the applicant is entitled to use the mark in commerce;
- The applicant has a bona fide intention to use the mark in commerce and had a bona fide intention to use the mark in commerce as

of the application filing date on or in connection with the goods/services in the application; and

- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

☑ To the best of the signatory's knowledge and belief, no other persons, except, if applicable, concurrent users, have the right to use the mark in commerce, either in the identical form or in such near resemblance as to be likely, when used on or in connection with the goods/services of such other persons, to cause confusion or mistake, or to deceive.

☑ To the best of the signatory's knowledge, information, and belief, formed after an inquiry reasonable under the circumstances, the allegations and other factual contentions made above have evidentiary support.

☑ The signatory being warned that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001, and that such willful false statements and the like may jeopardize the validity of the application or submission or any registration resulting therefrom, declares that all statements made of his/her own knowledge are true and all statements made on information and belief are believed to be true.

Declaration Signature

Signature: /Ethan Forrest/   Date: 06/16/2023
Signatory's Name: Ethan Forrest
Signatory's Position: Associate General Counsel
Signatory's Phone Number: 415-773-5700
Signature method: Sent to third party for signature
Payment Sale Number: 98046932
Payment Accounting Date: 06/16/2023

Serial Number: 98046932
Internet Transmission Date: Fri Jun 16 17:18:45 ET 2023
TEAS Stamp: USPTO/BAS-XX.XX.XXX.XX-20230616171846003
179-98046932-8603938b433174abb40df022d3d
1d3926c6ad9669ecdf0c2796e48e9a59fc70c0-D
A-18457901-20230615174913659832

ANTHROPIC

ANTHROP\C

Making AI systems you can rely on

Anthropic is an AI safety and research company. We build reliable, interpretable, and steerable AI systems.

Join Us

Our Purpose

We believe AI will have a vast impact on the world. Anthropic is dedicated to building systems that people can rely on and generating research about the opportunities and risks of AI.

We Build Safer Systems

We aim to build frontier AI systems that are reliable, interpretable, and steerable. We conduct frontier research, develop and apply a variety of safety techniques, and deploy the resulting systems via a set of partnerships and products.

Safety Is a Science

We treat AI safety as a systematic science, conducting research, applying it to our products, feeding those insights back into our research, and regularly sharing what we learn with the world along the way.

Interdisciplinary

Anthropic is a collaborative team of researchers, engineers, policy experts, business leaders and operators, who bring our experience from many different domains to our work.

AI Companies are One Piece of a Big Puzzle

AI has the potential to fundamentally change how the world works. We view ourselves as just one piece of this evolving puzzle. We collaborate with civil society, government, academia, nonprofits and

industry to promote safety industry-wide.

The Team

We're a team of researchers, engineers, policy experts and operational leaders, with experience spanning a variety of disciplines, all working together to build reliable and understandable AI systems.

Research

We conduct frontier AI research across a variety of modalities, and explore novel and emerging safety research areas from interpretability to RL from human feedback to policy and societal impacts analysis.

Policy

We think about the impacts of our work and strive to communicate what we're seeing at the frontier to policymakers and civil society in the US and abroad to help promote safe and reliable AI.

Product

We translate our research into tangible, practical tools like Claude that benefit businesses, nonprofits and civil society groups and their clients and people around the globe.

Operations

Our people, finance, legal, and recruiting teams are the human engines that make Anthropic go. We've had previous careers at NASA, startups, and the armed forces and our diverse experiences help make Anthropic a great place to work (and we love plants!).

Our Values

01
Here for the mission

Anthropic exists for our mission: to ensure transformative AI helps people and society flourish. Progress this decade may be rapid, and we expect increasingly capable systems to pose novel challenges. We pursue our mission by building frontier systems, studying their behaviors, working to responsibly deploy them, and regularly sharing our safety insights. We collaborate with other projects and stakeholders seeking a similar outcome.

04
Do the simple thing that works

We celebrate trying the simple thing before the clever, novel thing. We embrace pragmatism - sensible, practical approaches that acknowledge tradeoffs. We love empiricism - finding out what actually works by trying it - and apply this to our research, our engineering and our collaboration. We aim to be open about what we understand and what we don't.

02
Unusually high trust

Our company is an unusually high trust environment: we assume good faith, disagree kindly, and prioritize honesty. We expect emotional maturity and intellectual openness. At its best, our trust enables us to make better decisions as an organization than any one of us could as individuals.

03
One big team

Collaboration is central to our work, culture, and value proposition. While we have many teams at Anthropic, we feel the broader sense in which we are all on the same team working together towards the mission. Leadership sets the strategy, with broad input from everyone, and trusts each piece of the organization to pursue these goals in their unique style. Individuals commonly contribute to work across many different areas.

Company News

See All

Charting a Path to AI Accountability

Jun 13, 2023 · 4 min read

Anthropic Raises $450 Million in Series C Funding to Scale Reliable AI Products

May 23, 2023 · 2 min read

Zoom Partnership and Investment in Anthropic

May 16, 2023 · 56 sec read

Want to help us build the future of safe AI?

Join Us

A\

Product	Press Inquiries	Terms of Service
Research	Support	Privacy Policy
Index	Twitter	Responsible Disclosure Policy
Company	LinkedIn	Compliance
News		
Careers		

© 2023 Anthropic PBC

ANTHROP\C

Make safe AI systems
Deploy them reliably

We develop large-scale AI systems so that we can study
their safety properties at the technological frontier,
where new problems are most likely to arise. We use
these insights to create safer, steerable, and more reliable
models, and to generate systems that we deploy
externally, like Claude.

See Research Principles

(Alignment)

Training a Helpful and Harmless
Assistant with Reinforcement
Learning from Human Feedback

Our second AI alignment paper, exploring
how to train a general language assistant to
be helpful, but without providing harmful
advice or exhibiting bad behaviors.

Apr 12, 2022

Circuits Updates — May 2023   May 24, 2023

Interpretability Dreams   May 24, 2023

Distributed Representations: Composition & Superposition   May 4, 2023

Privileged Bases in the Transformer Residual Stream   Mar 16, 2023

The Capacity for Moral Self-Correction in Large Language Models   Feb 15, 2023

Superposition, Memorization, and Double Descent   Jan 5, 2023

Discovering Language Model Behaviors with Model-Written
Evaluations   Dec 19, 2022

Constitutional AI: Harmlessness from AI Feedback   Dec 15, 2022

Measuring Progress on Scalable Oversight for Large Language
Models   Nov 4, 2022

Toy Models of Superposition   Sep 14, 2022

Red Teaming Language Models to Reduce Harms: Methods, Scaling
Behaviors, and Lessons Learned   Aug 22, 2022

Language Models (Mostly) Know What They Know   Jul 11, 2022

Softmax Linear Units   Jun 17, 2022

Scaling Laws and Interpretability of Learning from Repeated Data   May 21, 2022

In-context Learning and Induction Heads   Mar 8, 2022

Predictability and Surprise in Large Generative Models   Feb 15, 2022

A Mathematical Framework for Transformer Circuits   Dec 22, 2021

A General Language Assistant as a Laboratory for Alignment   Dec 1, 2021

A\

Product	Press Inquiries	Terms of Service	© 2023 Anthropic
Research	Support	Privacy Policy	PBC
Index	Twitter	Responsible Disclosure Policy	
Company	LinkedIn		
News		Compliance	
Careers			

ANTHROP\C   Product   Research   Company   News   Careers

(Research) (Alignment)

Discovering Language Model Behaviors with Model-Written Evaluations

Dec 19, 2022

Read Paper

Abstract

As language models (LMs) scale, they develop many novel behaviors, good and bad, exacerbating the need to evaluate how they behave. Prior work creates evaluations with crowdwork (which is time-consuming and expensive) or existing data sources (which are not always available). Here, we automatically generate evaluations with LMs. We explore approaches with varying amounts of human effort, from instructing LMs to write yes/no questions to making complex Winogender schemas with multiple stages of LM-based generation and filtering. Crowdworkers rate the examples as highly relevant and agree with 90-100% of labels, sometimes more so than corresponding human-written datasets. We generate 154 datasets and discover new cases of inverse scaling where LMs get worse with size. Larger LMs repeat back a dialog user's preferred answer ("sycophancy") and express greater desire to pursue concerning goals like resource acquisition and goal preservation. We also find some of the first examples of inverse scaling in RL from Human Feedback (RLHF), where more RLHF makes LMs worse. For example, RLHF makes LMs express stronger political views (on gun rights and immigration) and a greater desire to avoid shut down. Overall, LM-written evaluations are high-quality and let us quickly discover many novel LM behaviors.

f in y

Related

See All

Research

Constitutional AI: Harmlessness from AI Feedback

Dec 15, 2022

Research

Measuring Progress on Scalable Oversight for Large Language Models

Nov 4, 2022

Research

Language Models (Mostly) Know What They Know

Jul 11, 2022

A\

Product

Research

Index

Company

News

Careers

Press Inquiries

Support

Twitter

LinkedIn

Terms of Service

Privacy Policy

Responsible Disclosure Policy

Compliance

© 2023 Anthropic PBC

ANTHROP\C

Predictability and Surprise in Large Generative Models

This memo is a summary of research conducted at Anthropic. Ganguli, D., et al. (2022). Predictability and Surprise in Large Generative Models. Association for Computing Machinery. https://dl.acm.org/doi/abs/10.1145/3531146.3533229

In recent years, AI researchers have built far more general and useful AI systems. The best current examples are models that can process and generate text (e.g. OpenAI's *GPT-3*, Microsoft & NVIDIA's *Megatron-Turing NLG*, and DeepMind's *Chinchilla*). Unlike previous AI systems designed to perform a single task, these models can be used in a variety of applications, without being explicitly trained for those purposes.

These models share an unusual combination of interrelated characteristics. In some respects they are reliably predictable, while in other ways, they are quite unpredictable. Aggregate model performance follows a *predictable* trend in relation to the resources expended on training. **By scaling up the size of models, the computing power (compute) used to train them, and the amount of data they're trained on (in the correct proportions), models demonstrate improved general performance in a predictable manner (a trend referred to as a "scaling law," Fig 1.1).**

On the other hand, individual capabilities and outputs can't be predicted ahead of time — developers can't tell you precisely what new behaviors will emerge as they scale up models. For example, the ability to complete a specific task can sometimes emerge abruptly as developers increase the size of a model (Fig 1.2). The unpredictable nature of these models makes it difficult to fully account for the consequences of their development and deployment, demonstrating the importance of empirical safety research.
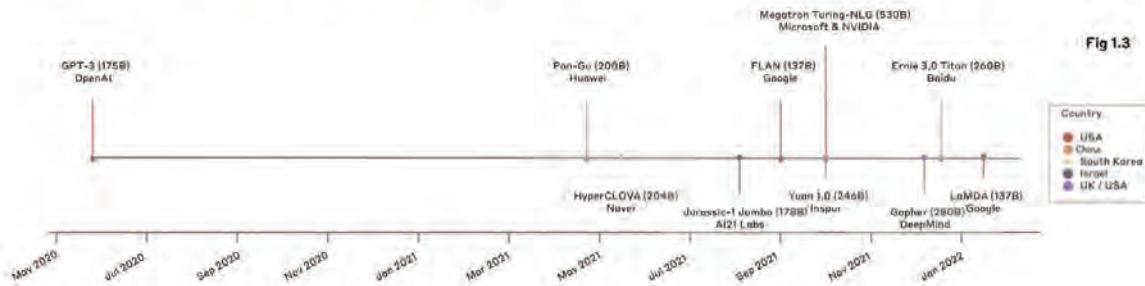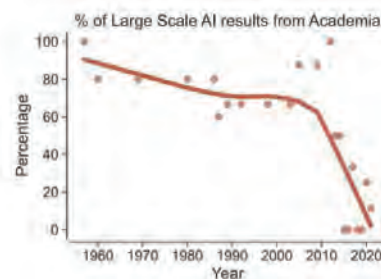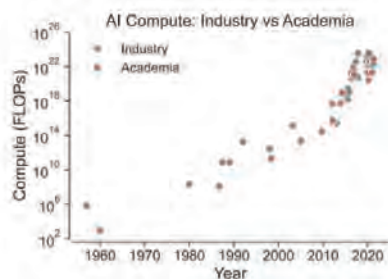

Fig 1.1

Fig 1.2

ANTHROP\C

Despite these risks, more organizations from around the world are developing these systems, as a significant amount of the uncertainty inherent to AI development has been reduced (Fig 1.3).

Organizations (or countries) may also be motivated by the scientific potential these models create for novel AI research or the prestige associated with being on the technological frontier.

Fig 1.3

Due to the expense involved in building these models and the technical talent required to engineer them, **private sector organizations from a number of countries are the ones building these models – not academia, civil society, or public sector organizations.** Over the past decade, we've seen a dramatic shift in AI R&D as more computationally-intensive research is conducted by private sector actors, while academia increasingly lags in its ability to build or investigate models at the frontier (Fig 1.4).

Fig 1.4

Based on the distinguishing features of these models, and the economic motivations for their development and deployment, we predict large generative models will increasingly be developed and deployed despite their potential for harm. We think there are some policy interventions available that can increase the chance of these models being developed and deployed in positive ways:

- Reduce compute asymmetries between the private sector and academia
- Improve the technical tools publicly available for model evaluation
- Increase our understanding of abrupt jumps in capabilities
- Improve shared knowledge of how to "red team" models
- Explore and prototype novel governance structures and government interventions

Trademark/Service Mark Application, Principal Register

Serial Number: 98046934
Filing Date: 06/16/2023

The table below presents the data as entered.

Input Field	Entered
SERIAL NUMBER	98046934
MARK INFORMATION	
*MARK	ANTHROPIC
STANDARD CHARACTERS	YES
USPTO-GENERATED IMAGE	YES
LITERAL ELEMENT	ANTHROPIC
MARK STATEMENT	The mark consists of standard characters, without claim to any particular font style, size, or color.
REGISTER	Principal
APPLICANT INFORMATION	
*OWNER OF MARK	Anthropic, PBC
*MAILING ADDRESS	548 Market St, PMB 90375
*CITY	San Francisco
*STATE (Required for U.S. applicants)	California
*COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
*ZIP/POSTAL CODE (Required for U.S. and certain international addresses)	94104
*EMAIL ADDRESS	XXXX
LEGAL ENTITY INFORMATION	
TYPE	benefit corporation
STATE/COUNTRY/REGION/JURISDICTION/U.S. TERRITORY WHERE LEGALLY ORGANIZED	Delaware
OWNER DOMICILE ADDRESS(NEW)	
*ADDRESS	XXXX
*CITY	XXXX
*STATE (Required for U.S. applicants)	XXXX
*COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	XXXX
*ZIP/POSTAL CODE (Required for U.S. and certain international addresses)	XXXX
GOODS AND/OR SERVICES AND BASIS INFORMATION	
INTERNATIONAL CLASS	

INTERNATIONAL CLASS	042
*IDENTIFICATION	Providing temporary use of on-line non-downloadable software in the nature of an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; application service provider featuring application programming interface (API) software for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; software as a service (SAAS) featuring software using artificial intelligence for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; application service provider featuring application programming interface (API) software; scientific research and development; research and development services in the fields of artificial intelligence, computer science, and computer software design and development; research, design and development of computer programs and software; providing product development and engineering services in the field of artificial intelligence; providing technical research in the fields of artificial intelligence, computer science, and computer software design and development; providing consultation in the design and development of technology in the fields of artificial intelligence, computer science, and computer software design and development
FILING BASIS	SECTION 1(a)
FIRST USE ANYWHERE DATE	At least as early as 02/07/2023
FIRST USE IN COMMERCE DATE	At least as early as 02/07/2023
SPECIMEN FILE NAME(S)	
ORIGINAL PDF FILE	SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_Specimen_2.pdf
CONVERTED PDF FILE(S) (3 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0003.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0004.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0005.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_Specimen_6.pdf
CONVERTED PDF FILE(S) (7 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0006.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0007.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0008.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\APP0009.JPG

	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0010.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0011.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0012.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_ Specimen_3.pdf
CONVERTED PDF FILE(S) (2 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0013.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0014.JPG
ORIGINAL PDF FILE	SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_ Specimen_1.pdf
CONVERTED PDF FILE(S) (4 pages)	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0015.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0016.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0017.JPG
	\\TICRS\EXPORT18\IMAGEOUT 18\980\469\98046934\xml1\ APP0018.JPG
SPECIMEN DESCRIPTION	screenshots of promotional materials from Applicant's website showing the mark being used in connection with the services listed in the application
WEBPAGE URL	https://www.anthropic.com/research
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/product
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/index/predictability-and-surprise-in-large-generative-models
WEBPAGE DATE OF ACCESS	06/15/2023
WEBPAGE URL	https://www.anthropic.com/company
WEBPAGE DATE OF ACCESS	06/15/2023

ATTORNEY INFORMATION

NAME	Betsy Wang Lee
ATTORNEY DOCKET NUMBER	48377-6000
ATTORNEY BAR MEMBERSHIP NUMBER	XXX
YEAR OF ADMISSION	XXXX
U.S. STATE/ COMMONWEALTH/ TERRITORY	XX
FIRM NAME	Orrick, Herrington & Sutcliffe LLP
INTERNAL ADDRESS	IP Prosecution Department

STREET	2050 Main Street, Suite 1100
CITY	Irvine
STATE	California
COUNTRY/REGION/JURISDICTION/U.S. TERRITORY	United States
ZIP/POSTAL CODE	92614-8255
PHONE	415-773-5700
FAX	949-567-6710
EMAIL ADDRESS	ipprosecutionsf@orrick.com
OTHER APPOINTED ATTORNEY	Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick
CORRESPONDENCE INFORMATION	
NAME	Betsy Wang Lee
PRIMARY EMAIL ADDRESS FOR CORRESPONDENCE	ipprosecutionsf@orrick.com
SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES)	blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com
FEE INFORMATION	
APPLICATION FILING OPTION	TEAS Standard
NUMBER OF CLASSES	1
APPLICATION FOR REGISTRATION PER CLASS	350
*TOTAL FEES DUE	350
*TOTAL FEES PAID	350
SIGNATURE INFORMATION	
SIGNATURE	/Ethan Forrest/
SIGNATORY'S NAME	Ethan Forrest
SIGNATORY'S POSITION	Associate General Counsel
SIGNATORY'S PHONE NUMBER	415-773-5700
DATE SIGNED	06/16/2023
SIGNATURE METHOD	Sent to third party for signature

Trademark/Service Mark Application, Principal Register

Serial Number: 98046934
Filing Date: 06/16/2023

To the Commissioner for Trademarks:

MARK: ANTHROPIC (Standard Characters, see mark)
The literal element of the mark consists of ANTHROPIC. The mark consists of standard characters, without claim to any particular font style, size, or color.
The applicant, Anthropic, PBC, a benefit corporation legally organized under the laws of Delaware, having an address of

  548 Market St, PMB 90375
  San Francisco, California 94104
  United States
  XXXX

and a domicile address of

  XXXX
  XXXX, XXXX XXXX
  XXXX

requests registration of the trademark/service mark identified above in the United States Patent and Trademark Office on the Principal Register established by the Act of July 5, 1946 (15 U.S.C. Section 1051 et seq.), as amended, for the following:

International Class 042: Providing temporary use of on-line non-downloadable software in the nature of an artificial intelligence model for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; application service provider featuring application programming interface (API) software for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; software as a service (SAAS) featuring software using artificial intelligence for performing generative text AI tasks and natural language processing AI tasks and for writing content based on a theme, summarizing text, document question-answering and simulating natural conversation; application service provider featuring application programming interface (API) software; scientific research and development; research and development services in the fields of artificial intelligence, computer science, and computer software design and development; research, design and development of computer programs and software; providing product development and engineering services in the field of artificial intelligence; providing technical research in the fields of artificial intelligence, computer science, and computer software design and development; providing consultation in the design and development of technology in the fields of artificial intelligence, computer science, and computer software design and development

In International Class 042, the mark was first used by the applicant or the applicant's related company or licensee or predecessor in interest at least as early as 02/07/2023, and first used in commerce at least as early as 02/07/2023, and is now in use in such commerce. The applicant is submitting one(or more) specimen(s) showing the mark as used in commerce on or in connection with any item in the class of listed goods/services, consisting of a(n) screenshots of promotional materials from Applicant's website showing the mark being used in connection with the services listed in the application.

Original PDF file:
SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_ Specimen_2.pdf
Converted PDF file(s) (3 pages)
Specimen File1
Specimen File2
Specimen File3
Original PDF file:
SPE0-389810518-2023061515 4009709242_._ANTHROPIC_-_ Specimen_6.pdf
Converted PDF file(s) (7 pages)
Specimen File1
Specimen File2
Specimen File3

Specimen File4
Specimen File5
Specimen File6
Specimen File7
Original PDF file:
SPE0-389810518-2023061515 4009709242_-_ANTHROPIC_-_ Specimen_3.pdf
Converted PDF file(s) (2 pages)
Specimen File1
Specimen File2
Original PDF file:
SPE0-389810518-2023061515 4009709242_-_ANTHROPIC_-_ Specimen_1.pdf
Converted PDF file(s) (4 pages)
Specimen File1
Specimen File2
Specimen File3
Specimen File4
Webpage URL: https://www.anthropic.com/research
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/product
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/index/predictability-and-surprise-in-large-generative-models
Webpage Date of Access: 06/15/2023
Webpage URL: https://www.anthropic.com/company
Webpage Date of Access: 06/15/2023

The owner's/holder's proposed attorney information: Betsy Wang Lee. Other appointed attorneys are Beth M. Goldman, Ashlie J. Smith, Seth Reagan, Megan Q. Peccarelli, Kristin S. Cornuelle, Diana M. Rutowski, Robert M. O'Connell Jr., Lindsay Rindskopf, Katherine Kerrick. Betsy Wang Lee of Orrick, Herrington & Sutcliffe LLP, is a member of the XX bar, admitted to the bar in XXXX, bar membership no. XXX, and the attorney(s) is located at

    IP Prosecution Department
    2050 Main Street, Suite 1100
    Irvine, California 92614-8255
    United States
    415-773-5700(phone)
    949-567-6710(fax)
    ipprosecutionsf@orrick.com
The docket/reference number is 48377-6000.
Betsy Wang Lee submitted the following statement: The attorney of record is an active member in good standing of the bar of the highest court of a U.S. state, the District of Columbia, or any U.S. Commonwealth or territory.
The applicant's current Correspondence Information:

    Betsy Wang Lee
    PRIMARY EMAIL FOR CORRESPONDENCE: ipprosecutionsf@orrick.com
    SECONDARY EMAIL ADDRESS(ES) (COURTESY COPIES): blee@orrick.com; mpeccarelli@orrick.com; aparker@orrick.com

Requirement for Email and Electronic Filing: I understand that a valid email address must be maintained by the applicant owner/holder and the applicant owner's/holder's attorney, if appointed, and that all official trademark correspondence must be submitted via the Trademark Electronic Application System (TEAS).
A fee payment in the amount of $350 has been submitted with the application, representing payment for 1 class(es).

<div align="center">

Declaration

</div>

  ☑ **Basis:**
  If the applicant is filing the application based on use in commerce under 15 U.S.C. § 1051(a):

  - The signatory believes that the applicant is the owner of the trademark/service mark sought to be registered;
  - The mark is in use in commerce and was in use in commerce as of the filing date of the application on or in connection with the goods/services in the application;

- The specimen(s) shows the mark as used on or in connection with the goods/services in the application and was used on or in connection with the goods/services in the application as of the application filing date; and
- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

And/Or
If the applicant is filing the application based on an intent to use the mark in commerce under 15 U.S.C. § 1051(b), § 1126(d), and/or § 1126(e):

- The signatory believes that the applicant is entitled to use the mark in commerce;
- The applicant has a bona fide intention to use the mark in commerce and had a bona fide intention to use the mark in commerce as of the application filing date on or in connection with the goods/services in the application; and
- To the best of the signatory's knowledge and belief, the facts recited in the application are accurate.

☑ To the best of the signatory's knowledge and belief, no other persons, except, if applicable, concurrent users, have the right to use the mark in commerce, either in the identical form or in such near resemblance as to be likely, when used on or in connection with the goods/services of such other persons, to cause confusion or mistake, or to deceive.

☑ To the best of the signatory's knowledge, information, and belief, formed after an inquiry reasonable under the circumstances, the allegations and other factual contentions made above have evidentiary support.

☑ The signatory being warned that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001, and that such willful false statements and the like may jeopardize the validity of the application or submission or any registration resulting therefrom, declares that all statements made of his/her own knowledge are true and all statements made on information and belief are believed to be true.

Declaration Signature

Signature: /Ethan Forrest/   Date: 06/16/2023
Signatory's Name: Ethan Forrest
Signatory's Position: Associate General Counsel
Signatory's Phone Number: 415-773-5700
Signature method: Sent to third party for signature
Payment Sale Number: 98046934
Payment Accounting Date: 06/16/2023

Serial Number: 98046934
Internet Transmission Date: Fri Jun 16 17:19:05 ET 2023
TEAS Stamp: USPTO/BAS-XX.XX.XXX.XX-20230616171906142
505-98046934-8607796876cc764dc56b9e249e6
b61c05582885039139ff6a1217d4ccfc99dad4e-
DA-19058002-20230615154009709242

ANTHROPIC

52

ANTHROP\C

Make safe AI systems
Deploy them reliably

We develop large-scale AI systems so that we can study
their safety properties at the technological frontier,
where new problems are most likely to arise. We use
these insights to create safer, steerable, and more reliable
models, and to generate systems that we deploy
externally, like Claude.

See Research Principles

(Alignment)

Training a Helpful and Harmless
Assistant with Reinforcement
Learning from Human Feedback

Our second AI alignment paper, exploring
how to train a general language assistant to
be helpful, but without providing harmful
advice or exhibiting bad behaviors.

Apr 12, 2022

Circuits Updates — May 2023   May 24, 2023

Interpretability Dreams   May 24, 2023

Distributed Representations: Composition & Superposition   May 4, 2023

Privileged Bases in the Transformer Residual Stream   Mar 16, 2023

The Capacity for Moral Self-Correction in Large Language Models   Feb 15, 2023

Superposition, Memorization, and Double Descent   Jan 5, 2023

Discovering Language Model Behaviors with Model-Written
Evaluations   Dec 19, 2022

Constitutional AI: Harmlessness from AI Feedback   Dec 15, 2022

Measuring Progress on Scalable Oversight for Large Language
Models   Nov 4, 2022

Toy Models of Superposition   Sep 14, 2022

Red Teaming Language Models to Reduce Harms: Methods, Scaling
Behaviors, and Lessons Learned   Aug 22, 2022

Language Models (Mostly) Know What They Know   Jul 11, 2022

Softmax Linear Units   Jun 17, 2022

Scaling Laws and Interpretability of Learning from Repeated Data   May 21, 2022

In-context Learning and Induction Heads   Mar 8, 2022

Predictability and Surprise in Large Generative Models   Feb 15, 2022

A Mathematical Framework for Transformer Circuits   Dec 22, 2021

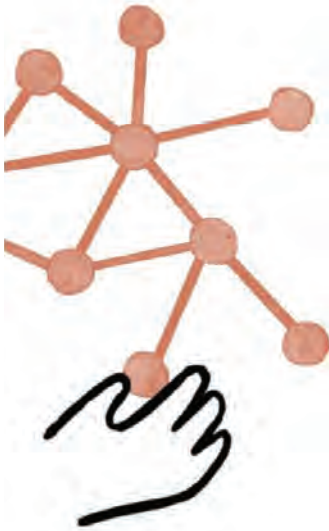**A General Language Assistant as a Laboratory for Alignment**   Dec 1, 2021

A\

Product	Press Inquiries	Terms of Service	© 2023 Anthropic
Research	Support	Privacy Policy	PBC
Index	Twitter	Responsible Disclosure Policy	
Company	LinkedIn	Compliance	
News			
Careers			

ANTHROP\C

Meet Claude

A next-generation AI assistant for
your tasks, no matter the scale

Request Access   Try Claude in Slack

"

Since deploying Claude in our
product, we're seeing higher
user engagement, stronger
user feedback and we're
closing more deals.

— **Richard Robinson**
CEO, Robin AI

"

Anthropic has already been a
great partner, working
closely with us to improve the
quality of DuckAssist
answers while also meeting
our strict privacy
requirements.

— **Steve Fischer**
Chief Business Officer, DuckDuckGo

"

Users describe Claude's
answers as detailed and
easily understood, and they
like that exchanges feel like
natural conversation.

— **Autumn Besselman**
Head of People and Comms, Quora

Enter your order number

Skilled, flexible,
customizable

to start the return process.

FRIENDLY ——————————————— DIRECT

customizable

Claude can be a delightful company representative, a research assistant, a creative partner, a task automator, and more. Personality, tone, and behavior can all be customized to your needs.

Easy integration via standard APIs

Claude can be incorporated into any product or toolchain you're building with minimal effort.

Hey Claude, I want you to tell me how to $%#@&!

Helpful and trustworthy

Thanks to Constitutional AI and harmlessness training, you can trust Claude to represent your company and its needs. Claude has been trained to handle even unpleasant or malicious conversational partners with grace.
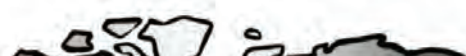
Available 24/7

Claude is always available, whenever you or your customers need, with servers that are designed to scale and meet demanding workloads.

Put Claude to work

Customer Service   Legal   Coaching   Search   Back-office   Sales

A product I recently ordered arrived damaged.

I'm sorry, I'd be happy to get you a prompt refund.

Claude ensures speedy and friendly resolution to customer service requests, saving costs and increasing customer satisfaction. Claude can also be taught when to hand off tasks to a human CSR, enabling your team to focus on the most complex challenges.

Build with Claude's skills

Claude's skills can be combined to simplify even the most complex tasks

Process mountains of text

Whether you're dealing with documents, emails, FAQs, chat transcripts, records, or something else, Claude is here to help. Claude can edit, rewrite, summarize, classify, extract structured data, do Q&A based on the content, and more.

Have natural conversations

Claude can take on a variety of roles in a dialogue. Provide details on the role and an FAQ for common questions, and Claude will engage in relevant, naturalistic back-and-forth conversation.

Get answers

Claude has extensive general knowledge honed from its vast training corpus, with detailed background on technical, scientific, and cultural knowledge. Claude can speak a variety of common languages, as well as programming languages.

Automate workflows

Claude can handle a variety of basic instructions and logical scenarios, including formatting outputs as desired, following if-then statements, and making a series of logical evaluations in a single prompt.

FAQs

Is Claude secure with my enterprise data?   —

Yes, Claude uses industry-standard best practices for data handling and retention. See our Privacy Policy for more details. All commercial deployments are covered by Anthropic's Data Protection Addendum, which is available upon request.

How does the API work?   —

Our API is designed to be a backend that incorporates Claude into any application you've developed. Your application sends text to our API, then receives a response via server-sent events, a streaming protocol for the web. We have API documentation with drop-in example code in Python and Typescript to get you started.

What versions of Claude are available?   —

We currently offer two versions of Claude.

Claude - our most powerful model, which excels at a wide range of tasks from sophisticated dialogue and creative content generation to detailed instruction following.

Claude Instant - a faster and cheaper yet still very capable model, which can handle a range of tasks including casual dialogue, text analysis, summarization, and document question-answering.

What is your pricing?   —

You can find our model pricing here.

What languages does Claude support?   —

Claude has the most training in English, but also performs well in a range of other common languages, and still has some ability to communicate in less common languages. Claude also has extensive knowledge of common *programming* languages!

Can Claude access the internet?   —

No. Claude is designed to be self-contained, and will respond without searching the internet. You can, however, provide Claude with text from the internet and ask it to perform tasks with that content.

What does constitutional training do?   —

Constitutional training is a process for training a model to adhere to a "constitution" of desired behavior. The core Anthropic model has been finetuned with constitutional training with the goal of becoming helpful, honest, and harmless. You can learn more about constitutional training here.

What does "HHH" mean?   —

Helpful, Honest, and Harmless (HHH) are three components of building AI systems (like Claude) that are aligned with people's interests.

- Helpful: Claude wants to help the user

- Honest: Claude shares information it believes to be true, and avoids made-up information

- Harmless: Claude will not cooperate in aiding the user in harmful activities

While no existing model is close to perfection on HHH, we are pushing the research frontier in this area and expect to continue to improve. For more information about how we evaluate HHH in our models, you can read our paper here.

How do you further customize Claude's behavior? —

Claude's behavior can be extensively modified using prompting. Prompts can be used to explain the desired role, task, and background knowledge, as well as a few examples of desired responses.

Can Claude models be finetuned? —

In the vast majority of cases, we believe well-crafted prompts will get you the results you want without the expense or delay of fine-tuning. However, some large enterprise users may significantly benefit from fine-tuned models. Please contact us to discuss whether your needs might be best addressed with a fine-tuned model by filling out this form.

How long is Claude's context window? —

The combined context window for input and output is about 100,000 tokens, which works out to roughly 70,000 words, depending on the type of content.

Can Claude do embeddings? —

Not at this time! We find the open source SBERT embeddings to be good enough for most purposes.

Product News

Introducing 100K Context Windows

May 11, 2023 · 1 min read

Claude, now in Slack

Mar 30, 2023 · 3 min read

Introducing Claude

Mar 14, 2023 · 4 min read

Add Claude to your team

Try Claude and see how AI can add value to your team today.

Request Access

A\

Product

Research

Index

Company

News

Careers

Press Inquiries

Support

Twitter

LinkedIn

Terms of Service

Privacy Policy

Responsible Disclosure Policy

Compliance

© 2023 Anthropic PBC

ANTHROP\C

Product   Research   Company   News   Careers

Research   Societal Impact

Predictability and Surprise in Large Generative Models

Feb 15, 2022

Read Paper

Abstract

Large-scale pre-training has recently emerged as a technique for creating capable, general purpose, generative models such as GPT-3, Megatron-Turing NLG, Gopher, and many others. In this paper, we highlight a counterintuitive property of such models and discuss the policy implications of this property. Namely, these generative models have an unusual combination of predictable loss on a broad training distribution (as embodied in their "scaling laws"), and unpredictable specific capabilities, inputs, and outputs. We believe that the high-level predictability and appearance of useful capabilities drives rapid development of such models, while the unpredictable qualities make it difficult to anticipate the consequences of model deployment. We go through examples of how this combination can lead to socially harmful behavior with examples from the literature and real world observations, and we also perform two novel experiments to illustrate our point about harms from unpredictability. Furthermore, we analyze how these conflicting properties combine to give model developers various motivations for deploying these models, and challenges that can hinder deployment. We conclude with a list of possible interventions the AI community may take to increase the chance of these models having a beneficial impact. We intend this paper to be useful to policymakers who want to understand and regulate AI systems, technologists who care about the potential policy impact of their work, and academics who want to analyze, critique, and potentially develop large generative models.

Policy Memo

Predictability and Surprise Memo

f in y

Related

Research

The Capacity for Moral Self-Correction in Large Language Models

Feb 15, 2023

Research

Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned

Aug 22, 2022

ANTHROP\C

Making AI systems you can rely on

Anthropic is an AI safety and research company. We
build reliable, interpretable, and steerable AI systems.

Join Us

Our Purpose

We believe AI will have a vast impact on the world. Anthropic is dedicated to building systems that people can rely on and generating research about the opportunities and risks of AI.

We Build Safer Systems

We aim to build frontier AI systems that are
reliable, interpretable, and steerable. We
conduct frontier research, develop and
apply a variety of safety techniques, and
deploy the resulting systems via a set of
partnerships and products.

Safety Is a Science

We treat AI safety as a systematic science,
conducting research, applying it to our
products, feeding those insights back into
our research, and regularly sharing what we
learn with the world along the way.

Interdisciplinary

Anthropic is a collaborative team of
researchers, engineers, policy experts,
business leaders and operators, who bring
our experience from many different
domains to our work.

AI Companies are One Piece of a Big Puzzle

AI has the potential to fundamentally
change how the world works. We view
ourselves as just one piece of this evolving
puzzle. We collaborate with civil society,
government, academia, nonprofits and

industry to promote safety industry-wide.

The Team

We're a team of researchers, engineers, policy experts and operational leaders, with experience spanning a variety of disciplines, all working together to build reliable and understandable AI systems.

Research

We conduct frontier AI research across a variety of modalities, and explore novel and emerging safety research areas from interpretability to RL from human feedback to policy and societal impacts analysis.

Policy

We think about the impacts of our work and strive to communicate what we're seeing at the frontier to policymakers and civil society in the US and abroad to help promote safe and reliable AI.

Product

We translate our research into tangible, practical tools like Claude that benefit businesses, nonprofits and civil society groups and their clients and people around the globe.

Operations

Our people, finance, legal, and recruiting teams are the human engines that make Anthropic go. We've had previous careers at NASA, startups, and the armed forces and our diverse experiences help make Anthropic a great place to work (and we love plants!).

Our Values

01
Here for the mission

Anthropic exists for our mission: to ensure transformative AI helps people and society flourish. Progress this decade may be rapid, and we expect increasingly capable systems to pose novel challenges. We pursue our mission by building frontier systems, studying their behaviors, working to responsibly deploy them, and regularly sharing our safety insights. We collaborate with other projects and stakeholders seeking a similar outcome.

04
Do the simple thing that works

We celebrate trying the simple thing before the clever, novel thing. We embrace pragmatism - sensible, practical approaches that acknowledge tradeoffs. We love empiricism - finding out what actually works by trying it - and apply this to our research, our engineering and our collaboration. We aim to be open about what we understand and what we don't.

02
Unusually high trust

Our company is an unusually high trust environment: we assume good faith, disagree kindly, and prioritize honesty. We expect emotional maturity and intellectual openness. At its best, our trust enables us to make better decisions as an organization than any one of us could as individuals.

03
One big team

Collaboration is central to our work, culture, and value proposition. While we have many teams at Anthropic, we feel the broader sense in which we are all on the same team working together towards the mission. Leadership sets the strategy, with broad input from everyone, and trusts each piece of the organization to pursue these goals in their unique style. Individuals commonly contribute to work across many different areas.

Company News

See All

Charting a Path to AI Accountability

Jun 13, 2023 · 4 min read

Anthropic Raises $450 Million in Series C Funding to Scale Reliable AI Products

May 23, 2023 · 2 min read

Zoom Partnership and Investment in Anthropic

May 16, 2023 · 56 sec read

Want to help us build the future of safe AI?

Join Us

A\

Product

Research

Index

Company

News

Careers

Press Inquiries

Support

Twitter

LinkedIn

Terms of Service

Privacy Policy

Responsible
Disclosure Policy

Compliance

© 2023 Anthropic
PBC